REPORT DOCUMENTATION PAGE	Form Approved OMB NO. 0704-0188

1. AGENCY USE ONLY (Leave Blank)	2. REPORT DATE February 3, 2006	3. REPORT TYPE AND DATES COVERED FINAL - May 16, 2005 - Nov. 15, 2005

4. TITLE AND SUBTITLE In Silico Screening for Biothreat Countermeasures	5. FUNDING NUMBERS W911NF-05-C-0063
6. AUTHOR(S) Lance M. Westerhoff, General Manager	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) QuantumBio Inc. 200 Innovation Blvd. Suite 261, State College, PA 16803	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) U. S. Army Research Office P.O. Box 12211 Research Triangle Park, NC 27709-2211	10. SPONSORING / MONITORING AGENCY REPORT NUMBER 49044.1-LS-CBD

11. SUPPLEMENTARY NOTES

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision, unless so designated by other documentation.

12 a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release; distribution unlimited.	12 b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

The current state of the art of in silico drug discovery or computer aided drug discovery relies almost exclusively on molecular mechanics force fields, such as AMBER and CHARMM, and empirical potentials. It is well known that while these approaches are excellent for certain applications, they have thus far proven less then satisfactory for thorough understanding the interactions of enzyme-inhibitor systems when used in Ki or IC50 prediction and best pose selection. To address this issue, in the first part of the proposed research, we will utilize our linear scaling, quantum mechanics algorithm and collaborate with our industrial partners to further research, develop, and validate a quantum mechanics-based score function, called QMScore, capable of predicting Ki and binding modes to the levels of accuracy required by the in silico drug discovery world. In the second part of the project, we utilize a knowledge management and artificial intelligence platform to aide in the usability of this advanced methodology by researching the relationships between structure and QM convergence – the ultimate goal being the development of an intelligent and adaptive system for drug discovery.

14. SUBJECT TERMS Linear-scaling quantum mechanics, protein-inhibitor interactions, in silico drug discovery, knowledge management, artificial intelligence, scoring function; computer aided drug discovery	15. NUMBER OF PAGES 19
	16. PRICE CODE

17. SECURITY CLASSIFICATION OR REPORT **UNCLASSIFIED**	18. SECURITY CLASSIFICATION ON THIS PAGE **UNCLASSIFIED**	19. SECURITY CLASSIFICATION OF ABSTRACT **UNCLASSIFIED**	20. LIMITATION OF ABSTRACT **UL**

Results of Phase I Work:

Phase I Proposal Abstract:

The current state of the art of in silico drug discovery or computer aided drug discovery relies almost exclusively on molecular mechanics force fields, such as AMBER and CHARMM, and empirical potentials. It is well known that while these approaches are excellent for certain applications, they have thus far proven less then satisfactory for thorough understanding the interactions of enzyme-inhibitor systems when used in Ki or IC50 prediction and best pose selection. To address this issue, in the first part of the proposed research, we will utilize our linear scaling, quantum mechanics algorithm and collaborate with our industrial partners to further research, develop, and validate a quantum mechanics-based score function, called QMScore, capable of predicting Ki and binding modes to the levels of accuracy required by the in silico drug discovery world. In the second part of the project, we utilize a knowledge management and artificial intelligence platform to aide in the usability of this advanced methodology by researching the relationships between structure and QM convergence – the ultimate goal being the development of an intelligent and adaptive system for drug discovery.

Phase I Aim 1 Progress: *QMScore Research, Development, and Validation"*

Previous to our Phase I effort, from an academic perspective, significant progress had been made with the research, and validation of QMScore for small sets of enzyme-inhibitor complexes[1, 2]. In the private sector, we have also worked with a number of pharmaceutical and biotech companies to further validate QMScore for use to supplement or perhaps even replace traditional empirical and molecular mechanics score functions. To date, due to non-disclosure agreements in place with the companies in question, none of these validations have been open and publishable. Therefore, as part of our research effort for the Phase I proposal, we further cultivated our relationship with Dr. David Diller at Pharmacopeia (see notes in "Key Personnel" section in Phase I proposal), and built an open and publishable QMScore set of general importance to our customers. Through this collaboration, we have further validated QMScore through execution of a project where QuantumBio will be permitted to publish the results of the "real world" validation. This validation set will act as a model to gage our success with aim 3 and most especially the Phase II effort. In addition to these practical applications of the set, we plan to publish the detailed results of this validation in a peer-reviewed journal.

For this validation run we chose two kinase p38 structures that are of significant interest to the drug discovery world. Dr. Diller worked to dock approximately 1200 ligands and drug candidates to each kinase structure using the well-known docking algorithm LibDock. This population of 1200 ligands includes ~400 ligands with experimentally quantified activity, ~400 with activity with other kinase structures, and ~400 with completely unknown activity. This resulted in a total of ~36,000 target/inhibitor complexes. We then took the resulting complexes and cleaned them up using the AMBER molecular mechanics force field as described in the Phase I proposal.

By July 25, 2005, the preparation of the structures for all 36,000 complexes was complete and we were ready to begin the quantum mechanical characterization of each of these complexes on the BlueGene/L supercomputer located at IBM's Capacity on Demand Center in Rochester, MN. Before performing these calculations however, the results from the AMBER force field calculations were utilized to help validate the final test set and to limit the number of redundant structures in the set. To do this, for each inhibitor, the AMBER energy was used to filter out energetically poor structures by introducing a 1.0Å clustering radius and filtering out those poses that did not meet the energetic requirements. Out of the remaining poses for each inhibitor, the top 10 poses, as measured by the AMBER energy, were kept as part of the test set. In this way, the quality of the test set is maintained, and this still left a total population of over 21,000 complexes to quantum mechanically characterize.

Initially, there were a number of problems associated with the early development of the BlueGene platform and a BlueGene-related bug or incompatibility in our source code. We were finally able to address these problems through continued development, and by calling upon a team of IBM developers and systems administrators to advise us on the tools available at our disposal on the platform. Once we were able to successfully execute entire scoring studies on BlueGene, the actual calculations ran extremely fast, and all 21,000 complexes were characterized within approximately five days on 1536 BlueGene nodes. This unprecedented level of performance provided us with an opportunity to study BlueGene in an actual run to better understand the strengths and weaknesses of the platform, and to better predict the requirements of our customers. For instance, while each BlueGene node is not as fast as a mainstream processor such as a Pentium or an Opteron, the computational density is extremely high. Further, once the calculations began, zero human interaction was required to complete the jobs. This latter characteristic is especially important as any calculation performed by a customer in an OnDemand capacity would need to be as self sufficient as possible.

Upon completion of this simulation, the populations of structures were considered, and the overall distribution of inhibitors was studied. In figures 1a and 1b, the output is provided for each of the targets (PDBid: 1ouy and 1a9u respectively) in the study complexed with its population of ~10,500 inhibitors. In the figure, the more negative the score the better. The three colored lines indicate the different inhibitor sets where the "P38 Score" data corresponds to the inhibitors of known activity (or inactivity) with that target, the "MDDR Score" data illustrates the score of the inhibitors generated through the MDDR approach, and finally the "Kinase Score" data is the score resulting from the non-P38 inhibitors that are inhibitors of other Kinase targets. By looking at the figure, it becomes clear that the QMScore did indeed correctly capture the relative interaction potential of the three groups. Further, by qualitatively considering the high scoring structures in the figure (those of Complex Number close to 1) from a conformational perspective, QMScore correctly determined poses that are similar to the known poses. This observation carries through even to the MDDR set and the Kinase set were in these cases, there are functional groups that share correct conformational and chemical attributes with the structures of known activity (ie: the P38 Set). These observations lead one to believe that, from a qualitative point of view, QMScore does correctly determine the best inhibitor in the correct conformation. Interestingly, for the 1a9u target/inhibitor set, QMScore had much more trouble distinguishing "known" actives from "unknown" actives. This result suggests that, as far as QMScore is concerned, the active site structure in 1a9u has taken on a conformation that is much less specific then the conformation found in 1ouy. For the target/inhibitor set at hand, a number of comparison studies are currently underway to consider how well competing scoring functions fare against this same population of inhibitors. The results of this comparison will be provided as part of the research paper to be published in the coming months.

It is encouraging that QMScore is able to differentiate between the different classes of inhibitors utilized in the study, however this does not tell us how well QMScore did at capturing a rank order score of the known inhibitors. Thanks to Dr. David Diller, we had access to the IC_{50} data of over 400 of the known p38 inhibitors used in the set. Based on this IC_{50} data, the affinities of these p38 inhibitors range from -5.9 kcal/mol through 0.0 kcal/mol (ie: inactive). Using this data as a starting point, the QMScore-determined affinity was compared to the experimental affinity as seen in figures 2a and 2b. In the results for 1ouy, the R^2 value is 0.1799. In the results for 1a9u, the R^2 value is 0.0796. Comparing these two results with figures 1a and 1b, again QMScore seems to do a better job treating the 1ouy target versus the 1a9u target. When considering figures 1 and 2 together, it does seem that QMScore is capturing the general trends in affinity, however more study is required to determine overall performance. In addition to using QMScore, which could be confounded by one or more overpowering interactions in one or more complexes, the Phase II SBIR effort will utilize the pair-wise decomposition scheme recently published in the research group[3] to more rigorously study the individual interactions between each target/inhibitor-complex. From this research we will gain insight not only into this population of inhibitors, but also the general applicability of this method in building an interaction profile – which is of much greater interest to our customers as evidenced by the letters of support in our Phase II proposal.

Figure 1a: Overall Relationship Between Inhibitor Classes Using "Pure" QMScore Results (PDBid: 1ouy)

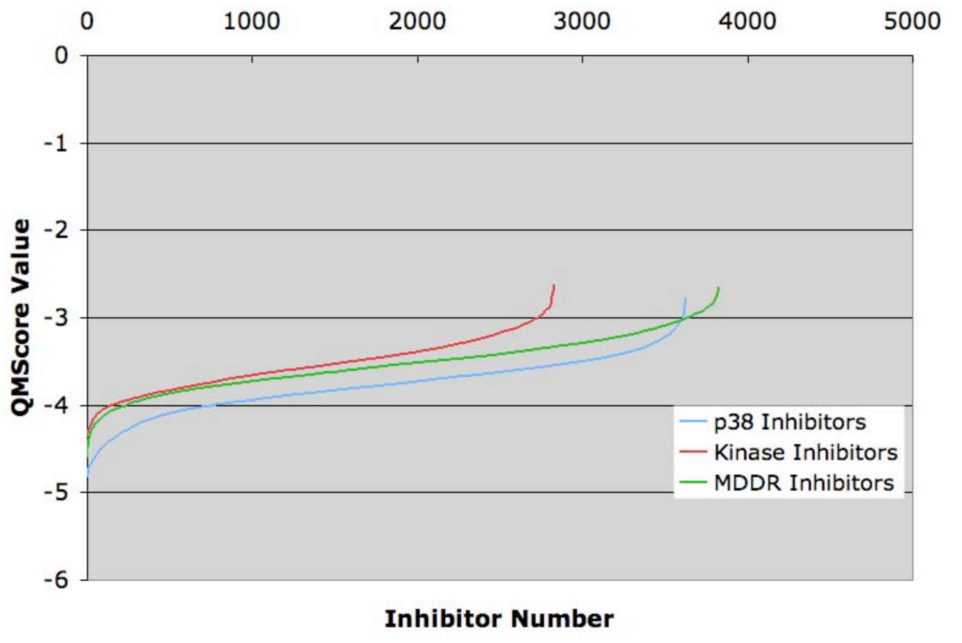


Figure 1b: Overall Relationship Between Inhibitor Classes Using "Pure" QMScore Results (PDBid: 1a9u)

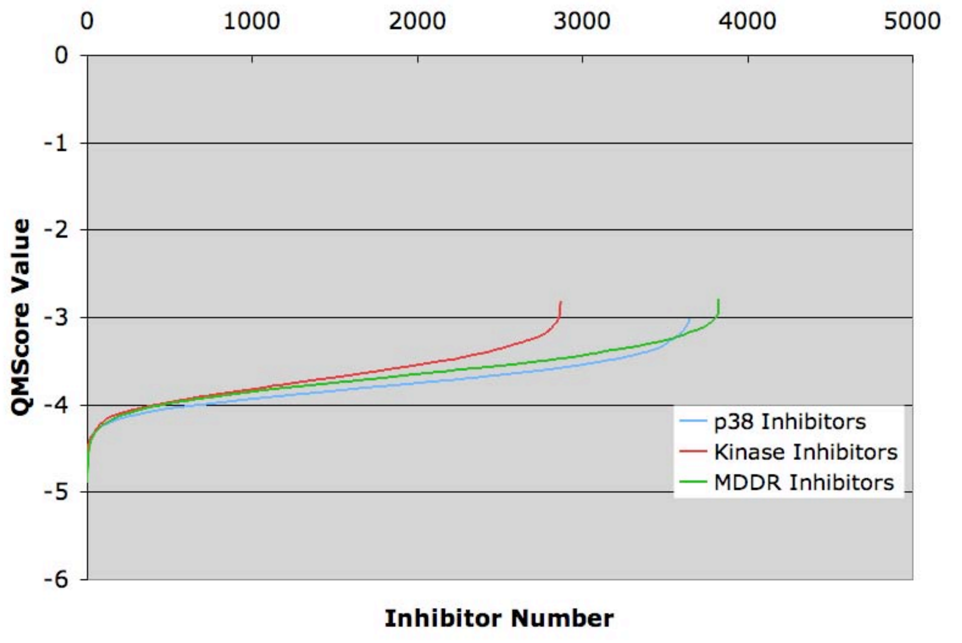


Figure 2a: Comparison Between Known Affinity and Predicted Affinity According to Pure QMScore (PDBid: 1ouy)

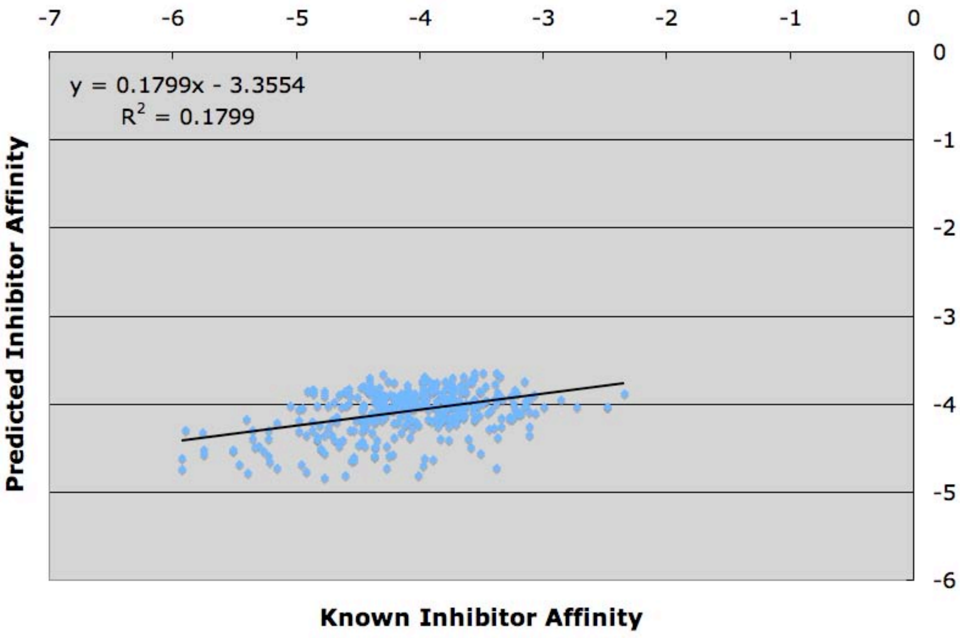
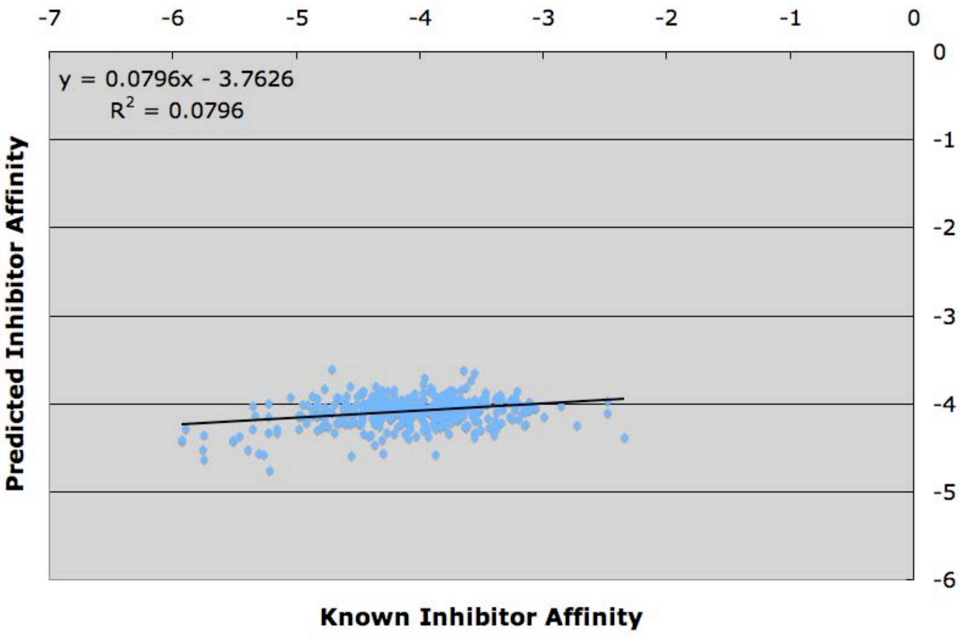


Figure 2b: Comparison Between Known Affinity and Predicted Affinity According to Pure QMScore (PDBid: 1a9u)



Another important goal of this research was to ascertain the relationship between the value of the score and the amount of conformational deviation (ie: root mean square deviation) exhibited in a particular inhibitor. As mentioned previously, each of the known p38 inhibitors in the population was docked to its target using the LibDock application. This method resulted in approximately 10 different energetically distinct poses for each p38 inhibitor. Many of these inhibitors share a core chemical structure with the inhibitors found in a x-ray crystal structures in the Protein DataBank (PDBids: 1ouy and 1a9u). Therefore, to understand how successful QMScore was in determining the "correct" conformation, a conformational analysis was performed on the inhibitors and compared to the score. The overall result of this analysis is shown in figures 3a and 3b in which a correlation is drawn between the predicted affinity (QMScore) and the root mean square deviation (RMSd) of the particular pose. All ~10 energetically dissimilar poses for each inhibitor were included in the analysis and the RMSd for the set of poses ranged from ~0.6Å to ~14Å. In general, when trying to determine the correct conformation based on score, the deviation (ie: RMSd) the better the score. As shown in figure 3a, for the 1ouy target, QMScore seems to be capturing that trend quite well in that one can observe that, in general the higher the score the lower the RMSd. This observation seems especially true if one were to consider the population in which the RMSd is ~4.0Å or less. As the red points in figure 3a denote, the regression is much more significant when only considering the inhibitors of the lowest root mean square deviation.

Since figure 3a considers all available poses for each inhibitor, the results can be further broken down to show that in over 23% of cases, the highest affinity conformer was also the one with the lowest RMSd. In fact, over 50% of cases, the highest affinity conformer had either the 1st or 2nd lowest RMSd. This translates into the highest affinity conformer having an RMSd of 3Å or better in over 50% of cases. Therefore, given a high quality (ie: low RMSd) conformation, QMScore's predictive ability is significantly improved. When another regression analysis was performed that considers only those poses that are within 1.5Å of the experimental conformation, the R^2 is 0.3785 compared to R^2 of 0.1799 shown in figure 2a. This result demonstrates that the quality of a QMScore prediction is indeed affected by the quality of a pose.

Figure 3a: Overall Comparison of QMScore and the RMSd Between the Inhibitor and a Known Conformation (PDBid: 1ouy)

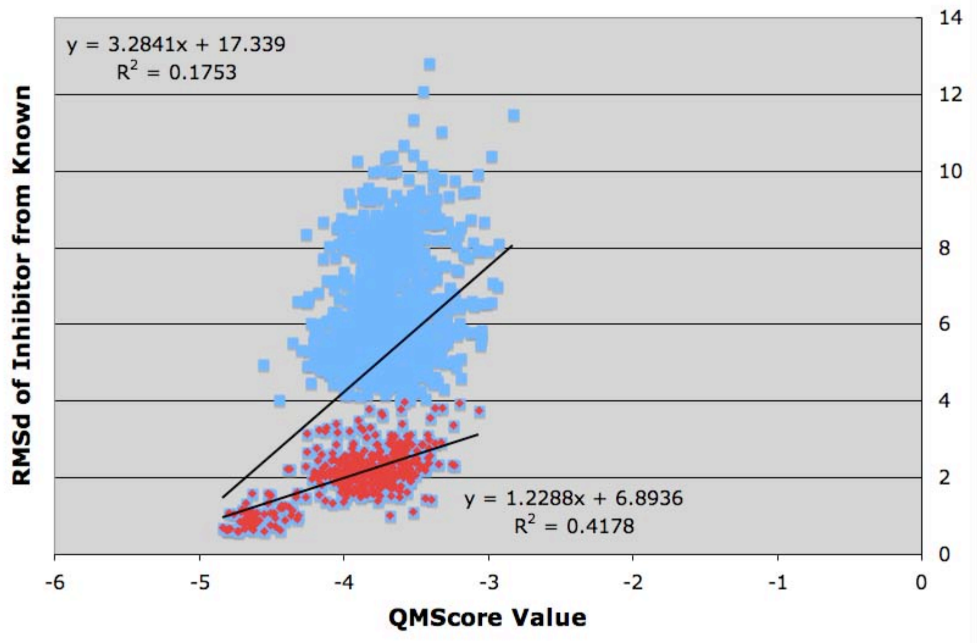
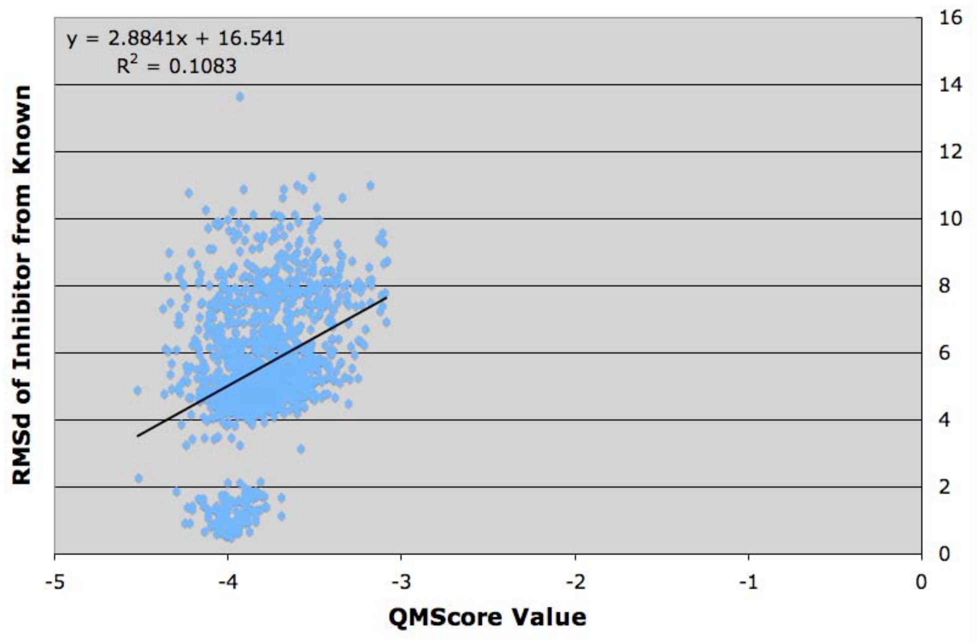


Figure 3b: Overall Comparison of QMScore and the RMSd Between the Inhibitor and a Known Conformation (PDBid: 1a9u)



MLR-Fitted Results

As noted in the literature[1, 2], each score value in QMScore is determined by summing together five key terms: the gas phase heat of formation, electrostatic solvation, attractive Lennard Jones, solvation entropy, and vibrational entropy. A pure QMScore calculation will utilize these terms without any coefficients. Past experience has shown that this pure QMScore analysis will capture most of the interaction between any given inhibitor with its target, and further fitting is not generally required to at least gain an overall understanding of performance. This benefit provides the user with a generalized scoring function that is applicable even to systems with limited experimental data. However, if one has a population of known inhibitors with known affinities, certainly one could utilize multiple linear regression (MLR) to determine a coefficient for each of the terms noted above and further train the score function for the structures at hand. When we performed the MLR analysis for the inhibitors of each target based on our set of known affinities resulting in a marked improvement in binding affinity prediction capabilities of the QMScore algorithm. Figures 4a and 4b show that the overall relationships between the different inhibitor classes are preserved and even accentuated when the MLR coefficients (see table 1) are utilized. However, QMScore still exhibits trouble differentiating between different inhibitor types in 1a9u, and this problem seems to permeate all aspects of our work with this target.

Table 1a: Multiple Linear Regression Coefficients (PDBid: 1ouy)

QMScore Term	Coefficient	Std. Error	T-Value	P-Value
Gas Phase ΔH	-0.0022297	0.0027147	-0.821	0.41199
Electrostatic Solvation	0.003515	0.0021065	1.669	0.09606
Attractive Lennard-Jones	-0.0106416	0.0033318	-3.194	0.00153
Solvation Entropy	0.0045091	0.0005987	7.531	4.07E-13
Vibrational Entropy [Ligand]	0.0296116	0.0139002	2.13	0.03382
Intercept	-2.0890241	0.2261274	-9.238	< 2e-16

Table 1b: Multiple Linear Regression Coefficients (PDBid: 1a9u)

QMScore Term	Coefficient	Std. Error	T-Value	P-Value
Gas Phase ΔH	0.0027519	0.0031666	0.869	0.385402
Electrostatic Solvation	0.007104	0.0034159	2.08	0.03826
Attractive Lennard-Jones	-0.0042565	0.0035012	-1.216	0.224878
Solvation Entropy	0.0026501	0.0007947	3.335	0.000943
Vibrational Entropy [Ligand]	0.0494679	0.0114619	4.316	2.06E-05
Intercept	-2.7346312	0.2766502	-9.885	< 2e-16

For the 1ouy target, the MLR analysis resulted in a set of coefficients that significantly improved the R^2 to 0.2238 (see figure 6a). Greater improvement can be reached by removing certain outlying points including all of those that are experimentally determined to be inactive and a number of particularly glaring points. The MLR coefficients are used throughout the RMSd analysis as well yielding improved results across the board (see figure 6 for MLR-Fitted versions of figure 3). This improvement is encouraging as it appears that in this case the MLR treatment does result in a better-trained QMScore model that is better able to distinguish between inhibitor classes, and rank order inhibitors.

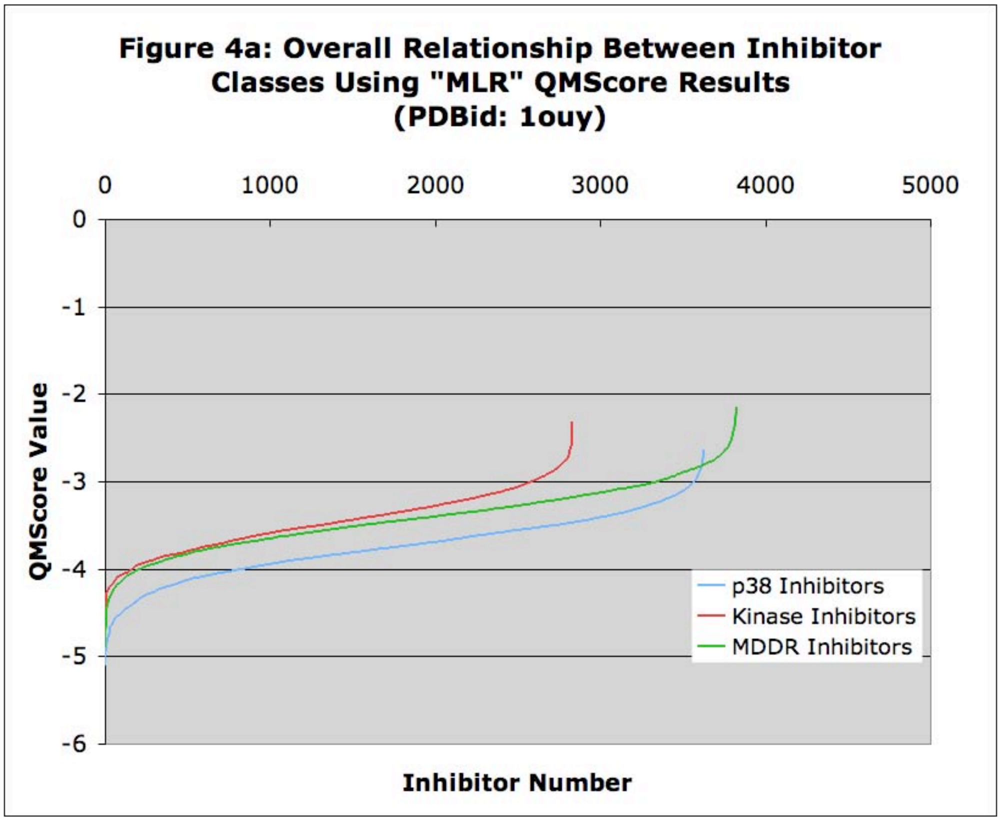


Figure 4a: Overall Relationship Between Inhibitor Classes Using "MLR" QMScore Results (PDBid: 1ouy)

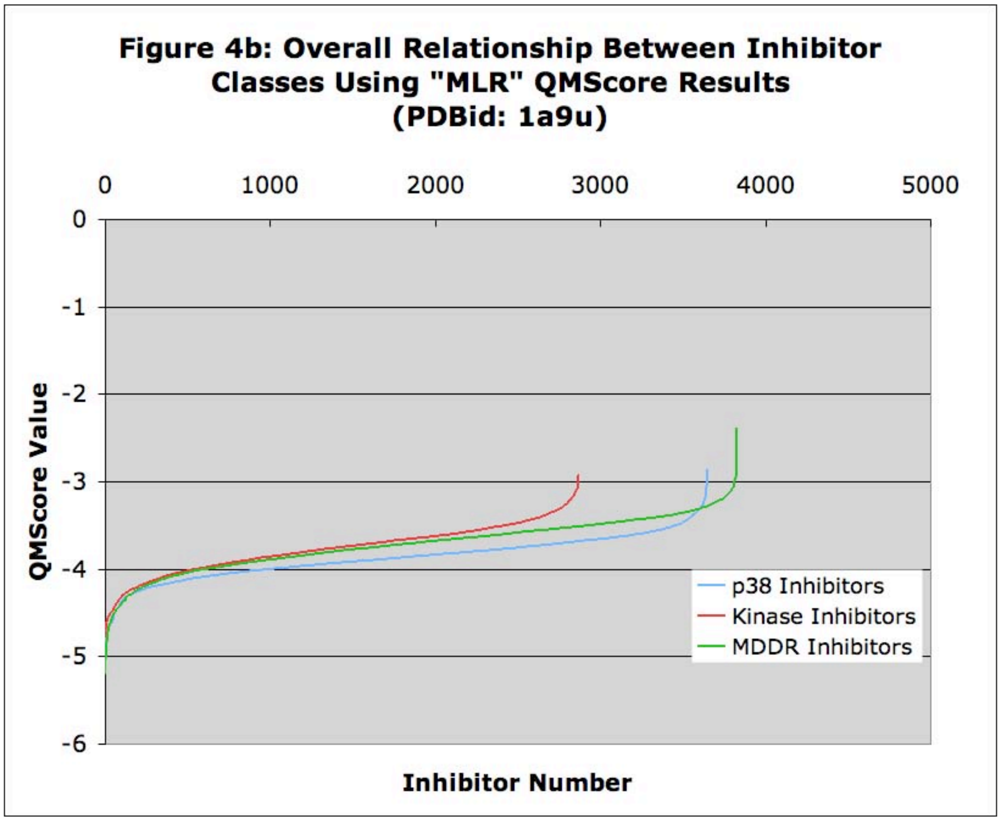


Figure 4b: Overall Relationship Between Inhibitor Classes Using "MLR" QMScore Results (PDBid: 1a9u)

Figure 5a: Comparison Between Known Affinity and Predicted Affinity According to MLR QMScore (PDBid: 1ouy)

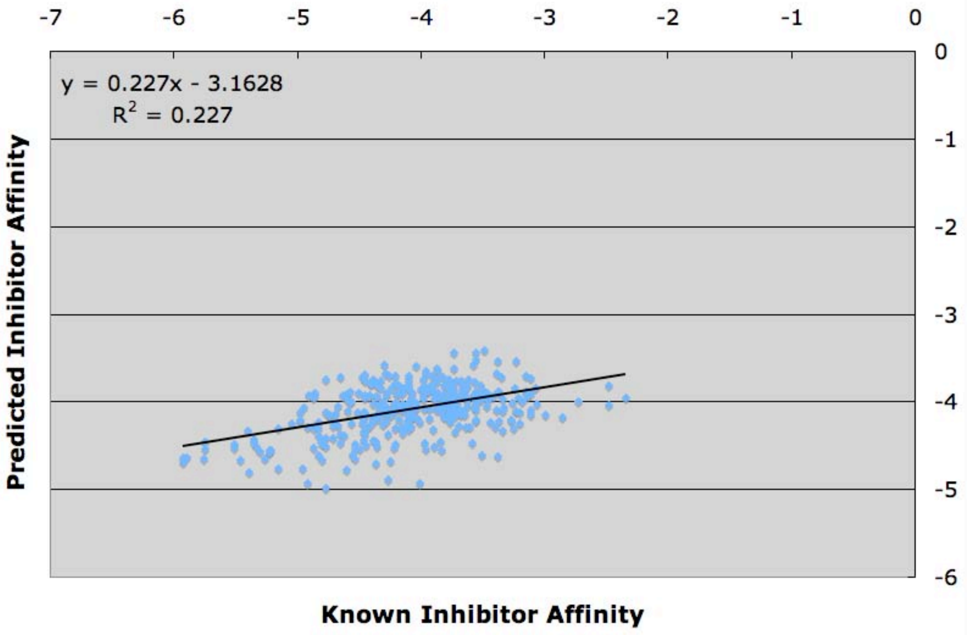


Figure 5b: Comparison Between Known Affinity and Predicted Affinity According to MLR QMScore (PDBid: 1a9u)

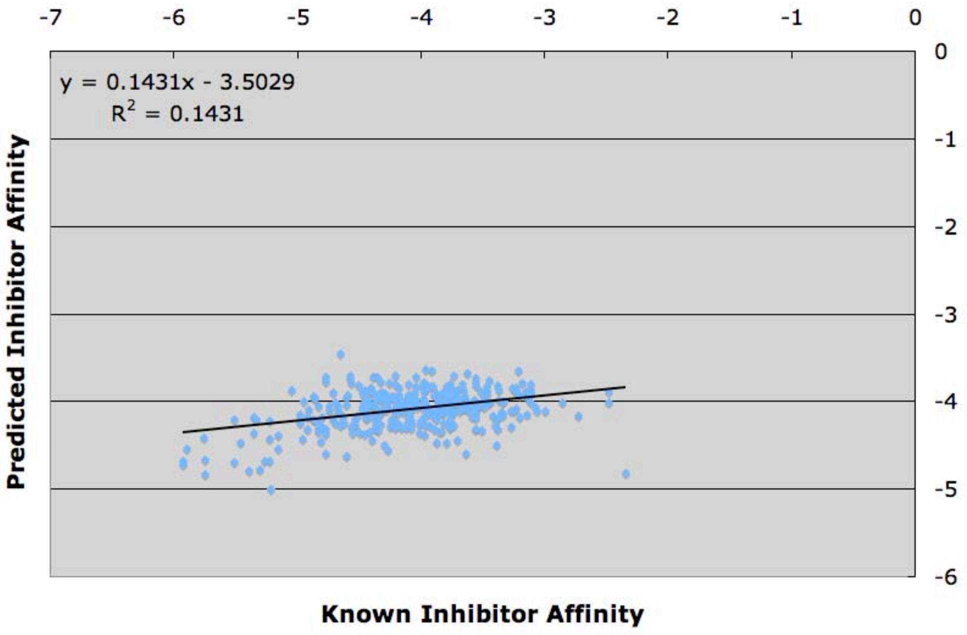


Figure 6a: Overall Comparison of MLR QMScore and the RMSd Between the Inhibitor and a Known Conformation (PDBid: 1ouy)

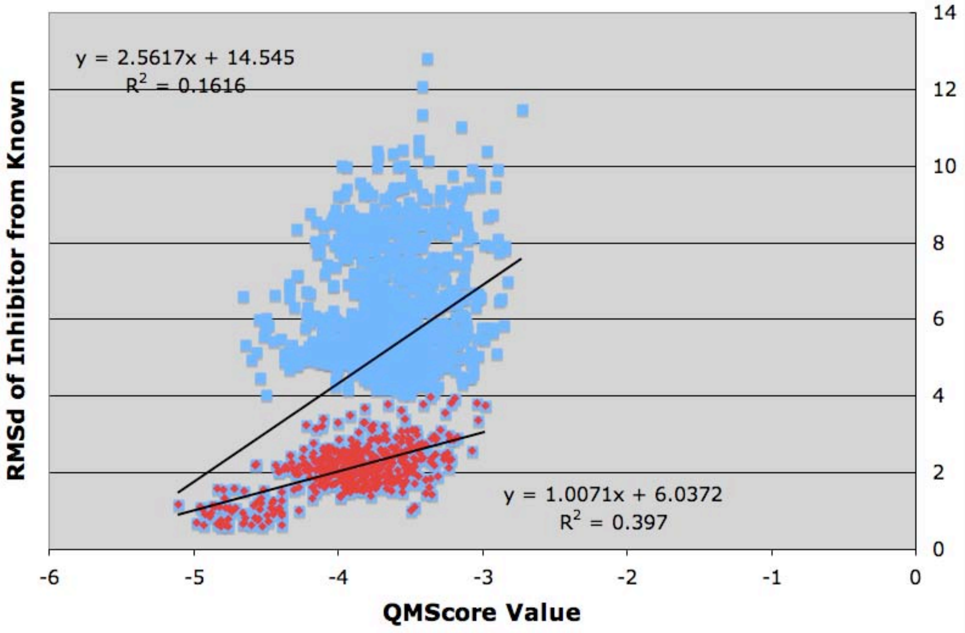
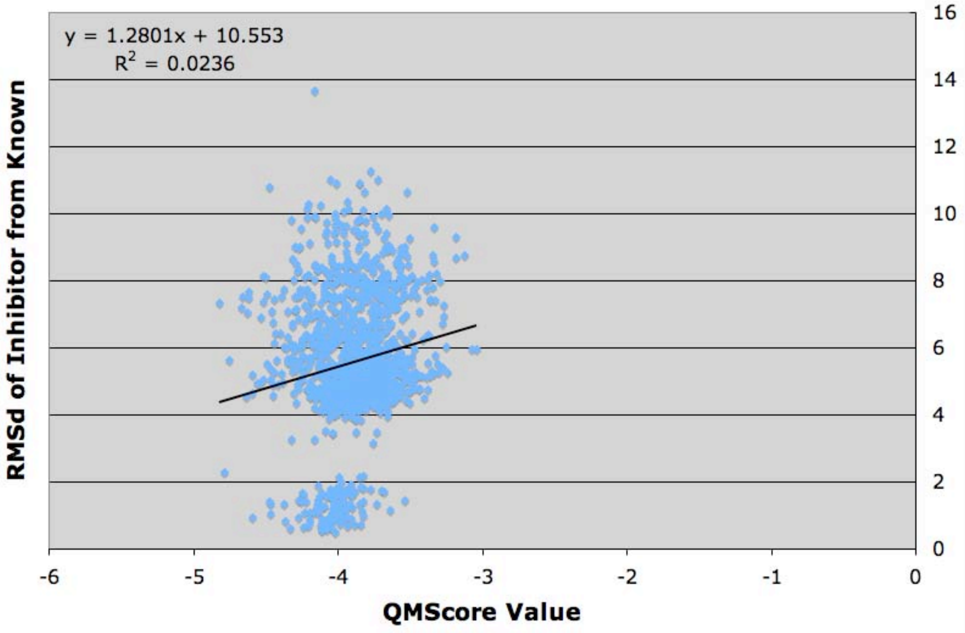
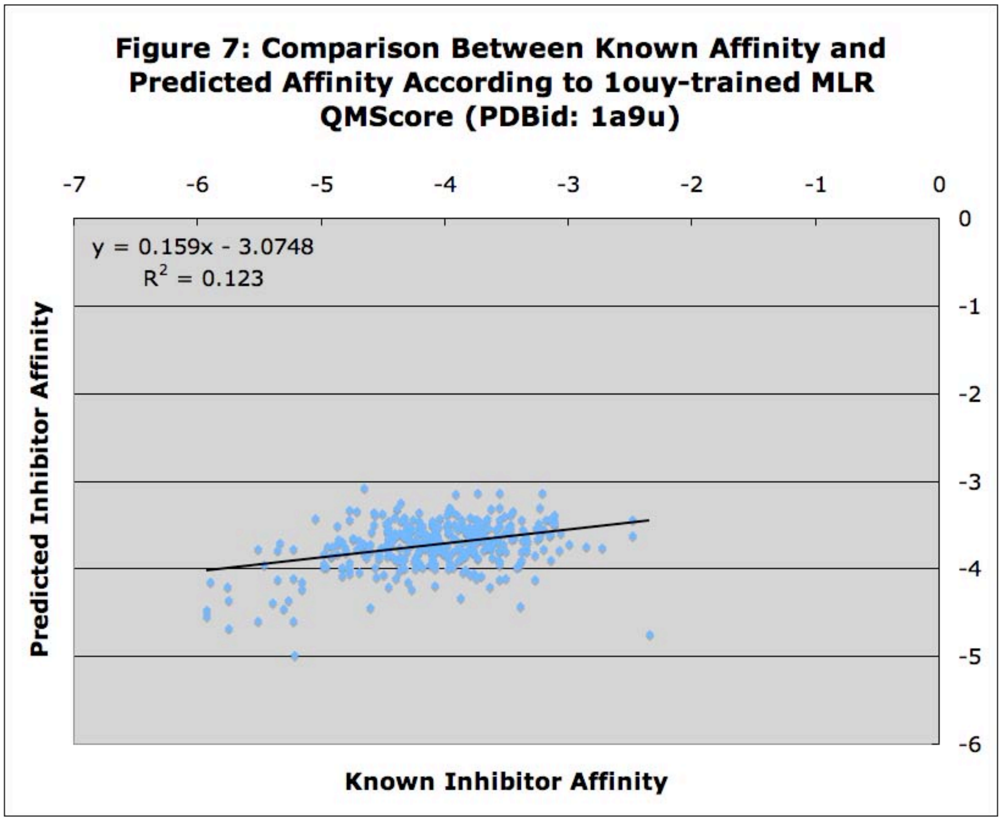


Figure 6b: Overall Comparison of MLR QMScore and the RMSd Between the Inhibitor and a Known Conformation (PDBid: 1a9u)



Initial Conclusions

In general, in the Phase I research aim, QMScore has been shown to perform adequately in distinguishing between different inhibitor classes – at least in the 1ouy case. More research is required to understand why QMScore performed so much better for the 1ouy case then the 1a9u case. The project has also shown that QMScore can be trained to yield a significantly improved model. With this understanding, it could be possible to build an automated workflow for our CHEMIX platform specifically to train the QMScore model based on a known set of inhibitors. This could have far reaching consequences as our customers utilize the software for any number of targets. It would also be possible to publish sets of coefficients for different target classes based on well-vetted training sets for each class. Figure 7 shows such an example where the MLR coefficients generated using 1ouy are applied to the 1a9u case. This particular example could be thought of as fairly contrived only because the 1a9u case is poorly understood, however it does illustrate that the coefficients could be used in such a fashion yielding an R^2 (0.123) that is still improved over the "pure" QMScore model (see figure 2b). Additional research is needed to determine how general this coefficient utility would be for other targets.



There are a number of analyses yet to be performed to fully appreciate the strengths and the weaknesses of the QMScore method. For one, it was hoped that QMScore would do an improved job in predicting the rank-order of the inhibitors. Certainly, trends can be shown, and if one were to go through and remove outliers, the correlation would be improved, however in the spirit of the project at hand – a real-life of application of the QMScore method – such over-processing would be counter-productive to truly understanding where we need to improve to develop a marketable product. Over the course of the next few weeks, various 3rd-party scoring functions will be utilized in order to compare our results to those of our competitors. It may become clear that our results are in agreement with or surpassing those of our peers. Ultimately, when we set out to validate QMScore, we purposefully chose a protein family that is known to be problematic simply to understand where we need to improve going forward. It is also hoped that the pair-wise decomposition (PWD) analysis will yield greater insights into the actual interactions that are taking place within each target-inhibitor complex ***regardless*** of the performance of the score function. Since score functions by nature consider the overall interactions between a target and its inhibitor in their totality, it is certainly logical that a PWD calculation, which explicitly treats each and every atom-atom interaction in a complex, will yield much greater insight even if the score itself captures very little. Based on repeated discussions with our colleagues, partners, and potential customers, this more complete treatment is actually of greater interest in the long run. Therefore, the Phase II effort will be centered on understanding and developing a united QMScore/PWD tool to be marketed to our customer base. It is expected that this so-called InteractionProfiler will aide the drug discovery world by providing our users with a much more complete appreciation of the important interactions found in each complex. The Phase II research effort will therefore leverage the validation performed in the Phase I by fully decomposing the interactions within the 1ouy and 1a9u test sets.

Phase I Aim 2 Report:

In recent years, we have created a molecular modeling platform, called CHEMIX, on which we will begin to build all subsequent software modules. This platform includes various analysis tools, molecular visualization widgets, scripting language support, and so on. As part of this Phase I SBIR, we built a preliminary link between this platform and the QMScore module as it stands today. In subsequent SBIR phases, we will further refine this link to build a commercially viable QMScore/PWD application. Aim 2 consisted of two parts:

1) The basic integration of QMScore with the CHEMIX platform.

2) Join the convergence and job characteristics data gleaned both during the Aim1 with various tools such as PROCHECK, WHATIF, and other external and internal software of this type to determine a population of pre-run, success/failure-descriptors

This latter aim was also in preparation for the Phase I (option) effort, so it required some expertise from our subcontractor.

Aim 2a "*Basic development of QMScore as a linkable module within our CHEMIX platform.*"

Thanks to the Phase I effort, CHEMIX now includes a QMScore module capable of aiding the user in the preparation of a QMScore job. This module was used to facilitate the rest of the project. As needed, the module will of course be refined to ultimately reach the goal of a marketable software tool. Included below are a number of "screenshots" of the tool in action. Each figure also shows an example of the graphical user interface (GUI) of the CHEMIX platform. Primarily, this interface consists of a Molecule Viewer, a Function Pallete, a Tool Bar, and various support windows (not shown). The "Plugin Dialog" window illustrates the GUI element of the QMScore module. As with many molecule viewers, the basic concept of a chemistry GUI is that the user is first presented with a "blank slate" into which he or she may import biological and chemical structures. These structures can be standard files found in publicly available databases, or in corporate databases. The GUI then allows the user to rotate, translate, or otherwise manipulate the imported structures. In the case of CHEMIX, the user may import any number of structures for use and manipulation. The user can then save a Project file that will include the CHEMIX-standard XML representation of each of these structures. This allows for easy organization and maintenance of a large number of biological and chemical structures. Once the user has a Project "in play," he or she may perform any number of simulations on the included structures. For the subject at hand, it is assumed that the user will run a QMScore calculation.

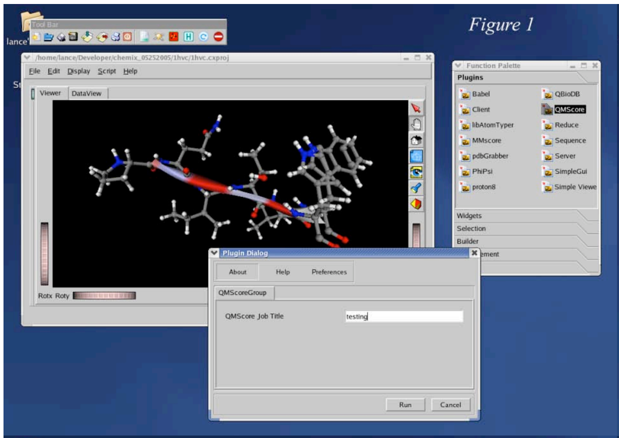


Figure 1

Plugin Dialog

QMScoreGroup

QMScore Job Title: testing

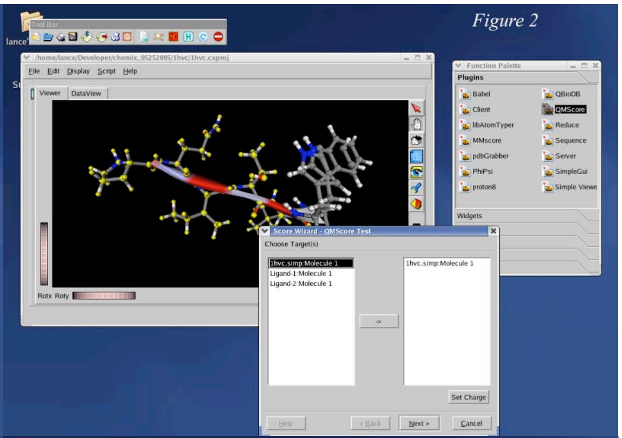


Figure 2

Score Wizard - QMScore Test

Choose Target(s)

1hvc.simp:Molecule 1
Ligand-1:Molecule 1
Ligand-2:Molecule 1

1hvc.simp:Molecule 1

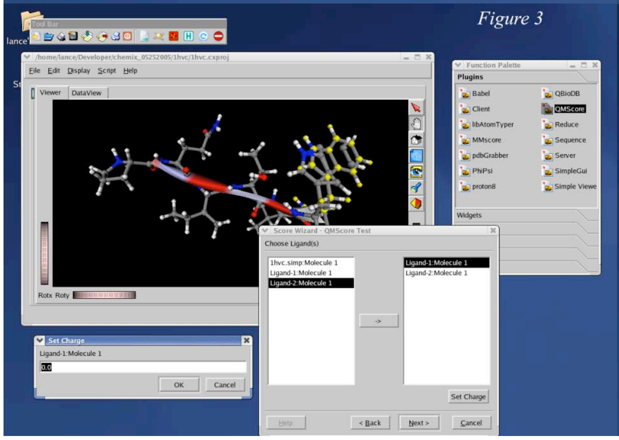


Figure 3

Score Wizard - QMScore Test

Choose Ligand(s)

1hvc.simp:Molecule 1
Ligand-1:Molecule 1
Ligand-2:Molecule 1

Ligand-1:Molecule 1
Ligand-2:Molecule 1

Set Charge

Ligand-1:Molecule 1

0.0

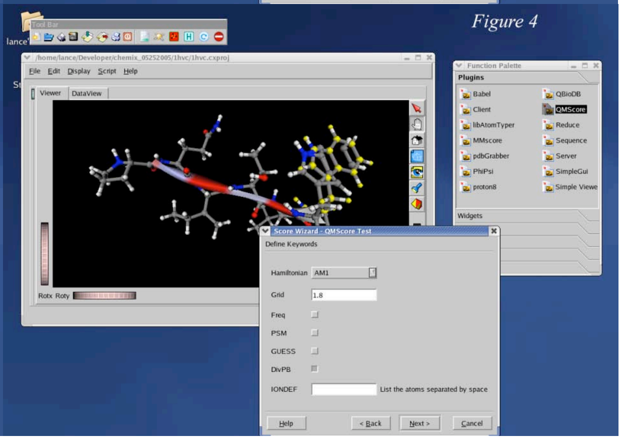


Figure 4

Score Wizard - QMScore Test

Define Keywords

Hamiltonian: AM1
Grid: 1.8
Freq ☐
PSM ☐
GUESS ☐
DivPB ☐
IONDEF: List the atoms separated by space

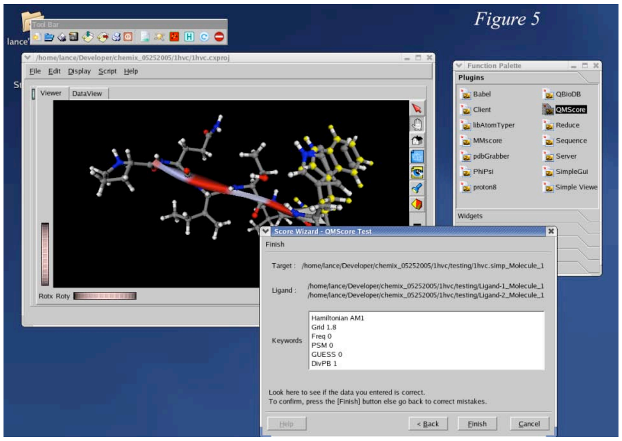


Figure 5

Finish

Target : /home/lance/Developer/chemix_05252005/1hvc/testing/1hvc.simp_Molecule_1

Ligand : /home/lance/Developer/chemix_05252005/1hvc/testing/Ligand-1_Molecule_1
/home/lance/Developer/chemix_05252005/1hvc/testing/Ligand-2_Molecule_1

Keywords

```
Hamiltonian AM1
Grid 1.8
Freq 0
PSM 0
GUESS 0
DivPB 1
```

Look here to see if the data you entered is correct.
To confirm, press the [Finish] button else go back to correct mistakes.

The primary function of the module is to provide the user with a "wizard" with which to build a QMScore "job" or "run" based on the target and inhibitors provided by the user. In figure 1, the user is asked to initialize the job with a title and the user may simply initialize the job with any name he or she requires. In this example, the Job Title is simply "Testing." As seen in figure 1, for simplicity, this example is purely contrived as it consists of a short polypeptide. The polypeptide is broken at its 6th residue and this last residue is replicated and shifted to provide the user with a 5-residue polypeptide followed by two separate small molecules. The polypeptide will simulate the "target" while the two separate small molecules represent two different inhibitors.

Once the user has chosen the name of the job, the next panel in the QMScore Wizard asks that the user define the "target" molecule. As illustrated in figure 2, the user is presented with a list of the molecules defined in the current CHEMIX project. In this case, the user has imported molecules named 1hvc.simp, Ligand-1, and Ligand-2. When the user selects each of these molecules in the list, the corresponding molecule in the Molecule Viewer will likewise be selected (as evidenced by the yellow spheres in the Viewer). This functionality acts to remind the user of which molecules to treat as "target" and which to treat as "ligand" by showing the user which item he has selected in the Target Chooser. This further underscores the level of integration between the CHEMIX platform and the QMScore module where selections and changes made in one directly affect the state of the other.

Figure 3, which shows the user choosing the ligands, illustrates functionality analogous to the Target Chooser in Figure 2. Since the current version of QMScore does require that the user input a total charge of each target and ligand in order to function correctly, a "Set Charge" button is also provided for the user. Using this mechanism, along with the associated "Set Charge" dialog box (also shown), the user can select each ligand and then set it corresponding total charge. This charge will then be stored with the molecule and used in the QMScore job as part of the simulation.

QMScore has a number of settings or keywords available to the user as shown in Figure 4. For example, the "Freq" keyword allows the user to run a frequency calculation on each inhibitor of the simulation. This is important for certain aspects of the simulation. The "Hamiltonian" provides the user with a list of available Quantum Mechanics Hamiltonians to be used in the QMScore calculation. Each of these items or keywords corresponds to those in the available documentation. It is this panel that will grow and change most as the project evolves going forward. Once the user has finished setting up their QMScore job, they would then click "Next" one final time.

Finally, the user is presented with a "summary" panel. As shown here, in this example, the user has opted to run a DivPB calculation using the AM1 Hamiltonian. The DivPB (or continuum solvent model) calculation will employ a grid size of 1.8. The "Target" input files along with the "Ligand" input files are shown at the top of the panel along with their complete path. Again, in order to improve user friendliness, this panel will evolve along with the panel shown in Figure 4, however the current version does illustrate the basic premise on which the QMScore setup module is built. In this final panel, once the user presses the "Finish" button, the module will save the correct input files for each of the structures chosen from the CHEMIX project along with the required run script. This run script can then be executed either locally on the users personal computer or remotely by using the tools to be developed in the Phase II effort.

Aim 2b – preparation *"Aim 3 (Phase I Option): Knowledge Management and Development."*:

The goal of Aim 2b was to characterize (using various 3rd party tools along with internal codes) the targets used in Aim 1 and record a number of descriptors both of the structure –bond lengths, bond angles, bond torsions, etc – and of the calculation –SCF counts, shift values, etc. However, it was found that in most cases, convergence problems were due to the target structure. This is logical in that in most instances in nature, and in all cases in the study, the target enzyme is much larger then the inhibitor. Therefore, with only two target molecules to consider, this population was not large enough to gain any useful insight into the causes of convergence problems in biomolecules. With this in mind, we expanded the population to approximately 3000 different protein structures from the Protein Data Bank. Each of these structures were prepared in much the same way as was denoted in the Phase I Work Plan for Aim 1, and they were all characterized using our DivCon application. In this case, the population was much larger, and we were more able to consider descriptors to provide us with insight into the empirical workings of the calculations used in QMScore. Armed with this knowledge, we will be better prepared to build the intelligent workflow system required to run the QMScore of the future. Our noted subcontractor – DiscoveryMachine – played a pivotal role in this process as it has yielded a statistical study of which characteristics are more important then others to determining.

DiscoveryMachine Inc. Results (for Phase I Aim 2b):

QMScore is a many-step process incorporating everything from protonation, atom typing, classical structure minimization and cleanup, to several steps of quantum mechanical characterization, environmental effects, and analytical tools. For any biomolecular system, any one of these steps could exhibit problems that will require the user to execute a contingency. To gain an understanding of how the structures prepared in each of these steps impacts convergence when running QMScore, DMI has begun development of a convergence analysis model comparing structure to convergence leveraging the recent QMScore runs.

The convergence analysis model is currently based on a multiple linear regression. DMI developed the analysis software application based on the statistical library JMSL, by Visual Numerics, an industry leader in mathematics and statistics software. The software developed by DMI allows the analysis to be automated and performed in a matter of seconds. Furthermore, the analysis software can be directly accessed by DMI strategy models via its Foreign Function Interface, enabling an expert to capture strategies for utilizing the statistical results. The current results are promising: out of over ten structures, only a handful seem to be important predictors of convergence.

Results of Analysis
The multiple linear regression analysis was performed on a large sample of structures (N = 2741) based on the number of iterations for convergence as the dependent variable and 12 predictors from the "What If" analysis. The sample set included samples in which convergence did not occur in 100 iterations; therefore, the range of iterations was 1 to 100 inclusive. The mean number of iterations in the sample was 54.1 (Std. Dev. = 27.6). The regression equation was highly significant (F = 49.4, df = 2728, p < 0.0001) with an R-squared value of 0.18. The following table shows the results for each predictor:

Predictor	Coefficient	T	P-value	Std. Error
Atom Count***	0.0022	5.0557	<< 0.0	0.0004
Residue Count***	0.0184	4.2279	<< 0.0	0.0044
First Gen Pack**	1.5639	2.982	0.0029	0.5244
Second Gen Pack	-1.2365	-1.7645	0.0778	0.7008
Rama Plot	4.3606	0.9617	0.3363	4.5343
Chi chi Rotamer	-1.8732	-0.4363	0.6627	4.2934
Backbone Conf***	1.9197	3.8821	0.0001	0.4945
Bond Lengths**	0.7952	2.7242	0.0065	0.2919
Bond Angles*	-3.7503	-2.3731	0.0177	1.5804
Omega Restraints	-0.0467	-0.0279	0.9778	1.676
Side Chain Plane	0.2788	0.3744	0.7081	0.7446
In Out Dist***	51.0522	6.0263	0	8.4716

* p <= .05
** p <= .01
*** p <= .001

The regression equation obtained was tested against 300 new samples. The correlation of the predicted number of iterations to the actual number of iterations for the new sample was 0.46. Further investigation was done into alternative types of regression models.

Other Commercialization Progress:

While we were working with the scientists at IBM to validate our software on their new platform, we were also fielding interviews with various news agencies including both Network World and InformationWeek in order to position our joint IBM-QuantumBio OnDemand offering in the months to come. Please see the following articles for more information:

http://www.networkworld.com/supp/2005/ndc5/082205-on-demand.html
http://www.informationweek.com/story/showArticle.jhtml?articleID=170100733
http://news.zdnet.com/2100-9584_22-5918025.html
http://www.betanews.com/article/IBM_Blue_Gene_Gets_New_Applications/1130522748
http://www.marketwire.com/mw/release_html_b1?release_id=99555
http://news.zdnet.co.uk/hardware/servers/0,39020363,39234112,00.htm
http://www.internetnews.com/bus-news/article.php/3559856

Further interviews will be performed as needed in the next few months to publicize our BlueGene efforts. In addition to these 3rd party efforts, QuantumBio is also working with IBM to finalize a press release outlining the availability of our software on BlueGene for customer use, and a research paper will be written in the coming months to discuss the results that were obtained during this QMScore validation.

Next Steps (Phase II Abstract):

The current state of the art of *in silico* drug discovery relies almost exclusively on molecular mechanics force fields, such as AMBER, and empirical potentials. It is well known that while these approaches are excellent for certain applications, they have thus far proven less then satisfactory for thorough understanding the interactions of enzyme-inhibitor systems. To address these issues, our linear scaling, quantum mechanics algorithm has been applied in the Phase I effort to further research, develop, and validate a QM-based score function, called QMScore. This score function showed itself to be capable of predicting Ki and binding modes to the levels of accuracy required by the *in silico* drug discovery world. In the Phase II, we will expand the validation performed in the Phase I, and leverage new industry collaborations, to demonstrate the power of our pair-wise decomposition algorithm when used in conjunction with QMScore. We will also develop the client/server software and database necessary to properly exploit these powerful QM tools in an industrial or a government setting. Finally, we will expand on the work performed in the Phase I that utilized an adaptive workflow environment that will yield an intelligent and adaptive system for drug discovery.

References

1. Raha, K. and K.M. Merz Jr., *A Quantum Mechanics Based Scoring Function: Study of Zinc-ion Mediated Ligand Binding*. J. Am. Chem. Soc., 2004. **126**: p. 1020-1021.
2. Raha, K. and K.M. Merz Jr., *Large Scale Validation of a Quantum Mechanics Based Scoring Function: Predictiing the Binding Affinity and the Binding Mode of a Diverse set of Protein-Ligand Complexes*. J. Med. Chem., 2005. **48**: p. 4558-4575.
3. Raha, K., A. van der Vaart, K.E. Riley, M.B. Peters, L.M. Westerhoff, H. Kim, and K.M. Merz Jr., *Pairwise Decomposition of Residue Interaction Energies Using Semiempirical Quantum Mechanical Methods in Studies of Protein-Ligand Interaction*. J. Am. Chem. Soc., 2005. **127**: p. 6583-6594.